AD641024

A MODEL FOR SCHEDULING MAINTENANCE UTILIZING MEASURES OF EQUIPMENT PERFORMANCE

ARINC RESEARCH MONOGRAPH NO. 8

C. E. Bradley, Ph.D., and E. L. Welker, Ph.D.

October 1, 1959

ARINC Research Corporation
(a subsidiary of Aeronautical Radio, Inc.)
1700 K Street, N.W.
Washington 6, D.C.

Publication No. 101-21-150

ENCL. (1) TO NAVSHIPS LTR SER 2021-T 77

Best
Available
Copy

ABSTRACT

This report describes a method for scheduling preventive maintenance to minimize expected average hourly maintenance cost based on a criterion of periodically observing deterioration in one or more equipment performance characteristics. The mathematical procedure requires expressing the deterioration phenomenon in the form of a simple Markov process. The implication of this method is that a forecast of equipment failure is based only on existing performance level and is independent of any history of prior deterioration rate. The criterion for scheduling preventive maintenance is expressed as a method involving matrix multiplication rather than as a simple algebraic formula or a series of curves. This was necessitated by the large number of input parameters consisting of maintenance cost parameters and a matrix of probabilities descriptive of the deterioration phenomenon.

Hypothetical numerical examples established the potential of this method for achieving real saving in maintenance cost. The method provides a systematic search for "lemon" equipments, and, conversely, protects against discarding those equipments which tend to maintain high performance levels over extended periods of time.

As an added result of this analysis, the algebraic method provides a technique for collecting deterioration data in terms of distributions and not just averages.

It was apparent in the numerical work that the underlying failure density function is critical in determining the amount of saving which can be achieved by this method. The coefficient of variation is especially critical. More theoretical studies and field data collection in this area are indicated. It is important to observe, however, that the method is distribution free since it does not depend on prior knowledge of the time to failure density function. However, the methods of data collection, definition of states or performance levels, and the selection of proper time intervals present peculiar problems which require care in the application of this method. These points are discussed in some detail in the Appendix.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

The first ARINC Research Corporation monograph in this series presented a method for determining a preventive maintenance schedule based upon part replacement prior to in-service failure.* The model used in that monograph related equipment operating time between such preventive maintenance actions to the expected average hourly cost of maintenance. This model was based on the assumption that a measure of equipment deterioration was not available; consequently, the method is applicable in those cases where deterioration is the primary cause of failure but cannot be measured. For those cases in which it is possible to monitor equipment performance periodically, and thus to measure the degree of deterioration, a different procedure for the selection of a minimum-cost preventive maintenance schedule can be used. The development of such a procedure is the purpose of this second paper in the series.

Basic to this paper are the assumptions that gradual deterioration is the principal cause of failure in many equipments, and that deterioration is reflected in the experimental data collected in an equipment study. Measurements of equipment deterioration, together with associated failure probabilities and cost parameters, are used to compute expected average hourly maintenance costs. The "optimum" preventive maintenance schedule is that in which the expected average

---

* Welker, E. L., Relationship Between Equipment Reliability, Preventive Maintenance Policy, and Operating Costs (Monograph No. 7), ARINC Research Corporation, Washington, D. C., February 13, 1959 (Publication No. 101-9-135).

hourly cost of maintenance is at a minimum. The answer is expressed in terms of the level of equipment performance and the time between performance measurements.

1.1 Type of Maintenance Situation Described in the Paper

The maintenance situation modeled here is illustrated by the case in which maintenance personnel for a fleet of trucks periodically measure the depth of tire tread. As the tread wears down there is an increasing probability of tire failure in a given subsequent time period. If failure of the tire in service is associated with costs above that of the actual cost of replacing the tire (e.g., a blown tire would result in lost man-hours or a wreck) it is desirable to replace the tire at some convenient scheduled time prior to the time of this in-service failure. Replacing the tire too soon will increase the operating cost through wasted tire-miles. Replacing the tire after the tread has become extremely worn will mean high cost through a high in-service failure rate. Obviously there is an "optimum" depth of tread which, from a cost standpoint, warrants replacement of the tire.

Similarly, the ratio of signal-plus-noise to noise in a radio receiver might be a performance characteristic upon which a computation of "optimum" time for repair could be based. As the ratio of signal-plus-noise to noise decreases, there is an increasing probability that the receiver will fail in operation within a given subsequent time period. Intuitively, there should be some optimum ratio of signal-plus-noise to noise which, from a cost standpoint, warrants equipment repair.

1.2 Critical Basic Assumptions

As indicated earlier, two assumptions are implicit in the method to be described. It is assumed, first, that system failure results primarily from a process of degeneration with time, and, second, that the most significant wear-out phenomena are reflected in the performance characteristic which is measured. Thus, changes in transconductance, an increase in leakage, and mistuning might all be reflected in, for example, the ratio of signal-plus-noise to noise. Obviously, it is rare to find all modes of failure reflected in a single measured characteristic. In the tire illustration, measurements of tread would give no direct clue to the probability of sidewall failure; and, if sidewall failure contributed significantly to in-service failure, measurement of tread alone would be an insufficient test. In the following material, only one measurement of equipment performance is discussed, although the method can be extended to those situations in which several measurements are taken (see Section 2.1.4).

1.3 Definition of Failure

Equipment failures are usually determined by either or both of two classes of personnel -- by maintenance personnel, who evaluate the operational level of the equipment through periodic tests or measurements; or by operating personnel, who evaluate the equipment through observation of performance, the end out-put of the equipment operation.

During periodic checks maintenance personnel will remove an equipment from service if in their judgment the performance level is too low to be adequate. Thus, the equipment may be removed from service even though the operator does not express dissatisfaction with

performance. On the other hand, the operator may reject the equipment even if its performance -- as subsequently measured by the maintenance man -- is fairly high. An equipment which is taken out of service for repair as a result of dissatisfaction on the part of either the maintenance man or the operator is defined as a failure in this paper. Maintenance actions performed as a result of such removal from service are to be distinguished from maintenance actions on equipments whose level of performance is considered adequate by both groups of personnel. The latter actions are defined as preventive maintenance.

1.4 The Maintenance Problem

The maintenance problem can be examined in more detail by reference to Figure 1. The curve, beginning with peak performance scaled at 1.0, is typical of the average level of a measured performance characteristic of an equipment over time. (See Section 2.2.2 for a discussion of variations about the curve.) The dotted line in Figure 1 is the level of the measured characteristic at which failure

AR-0203

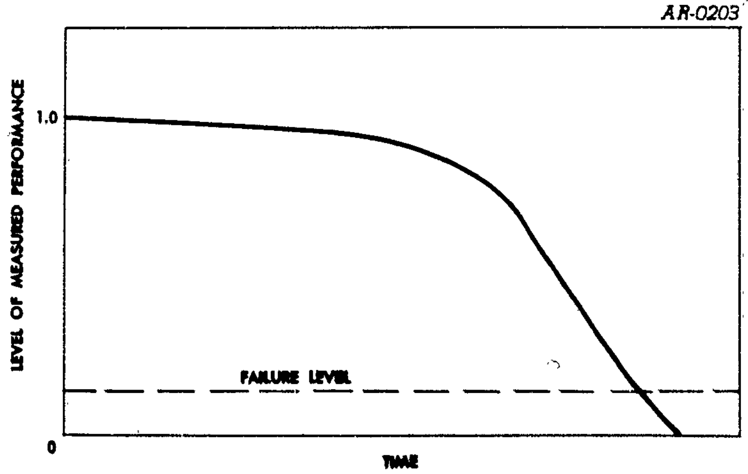

FIGURE 1
TYPICAL DETERIORATION OF A PERFORMANCE CHARACTERISTIC

occurs -- a level below which there is agreement that the performance of the set will not be satisfactory. It is recognized that there is no clear line of demarcation between satisfactory and unsatisfactory performance; the level indicated by the dotted line in Figure 1 is introduced only for reasons of exposition. Possibly it is better stated that there exists a different probability of failure -- as defined in Section 1.3 -- with each performance level. This difference in the probability of failure is reflected in the probability matrix to be used, and is discussed in detail later.

The maintenance personnel examine an equipment at time t and note the measured level of operation. If performance has a value comparatively close to the dotted line, the decision may be made to repair. This is particularly true if maintenance personnel conclude that there is a high probability of failure prior to the next maintenance inspection. The time interval to the next inspection is obviously involved, since the longer this interval the greater the probability of failure in the interval. Thus, the longer the time interval to the next inspection the higher must be the level of performance at which preventive action should take place. The question to be answered is: What level of performance with what time interval between inspections will provide minimum operating cost? The model developed in Section 2 is offered as one method by which the answer to this question may be determined.

2. GENERAL METHODOLOGY

2.1 The Markov Process

The maintenance situation described in the previous section implies a model which can be described in terms of a Markov Process.* In order to do this, it is useful to think of the equipment performance characteristics as discrete variables whose separate measured levels are called "states." Similarly, it is convenient to assume a discrete time variable for periodic performance measurement. This time interval can be associated with the concept of "trial" commonly encountered in discussions of Markov processes. In a Markov process, one is concerned with the probabilities of transition from one state to another in a single trial. The analogue in the present case is the probability of a transition from one equipment performance level to another in the time interval between inspections. It is now necessary to describe these probabilities in mathematical form, using the words "state" and "trial" for brevity of expression and to facilitate reference to the discussions of the Markov process in the literature.

The essential element of a Markov process is a set of conditional probabilities, p_{ij}, the probability that if the equipment is known to be in state i it will pass to state j in a single trial. These probabilities can be conveniently written as a matrix, called the

* For a discussion of discrete Markov chains, see Feller, W., An Introduction to Probability Theory and Its Applications, John Wiley & Sons, Inc., 1950, Chapters 15 and 16.

transition matrix of the Markov process.* For a simple case in which there were two states, the transition matrix would appear as:

$$\begin{array}{c c} & \begin{array}{cc} a_1 & a_2 \end{array} \\ \begin{array}{c} a_1 \\ a_2 \end{array} & \left\| \begin{array}{cc} p_{11} & p_{12} \\ p_{21} & p_{22} \end{array} \right\| \end{array} .$$

For clarity, the states are indicated in the rows and columns as a_1 and a_2. The matrix value p_{11} is the probability that if state a_1 existed at trial k, this state would still exist in the next trial, k + 1. The value p_{12} is the probability that if state a_1 existed at trial k, it would change to state a_2 in trial k + 1. Thus, we have a set of conditional probabilities: given that state a_i exists, p_{ij} is the probability of being in state a_j following the immediately subsequent trial. Since the above transition matrix states only what happens during the transition should state a_i exist, it is necessary to know initially what state does, in fact, exist.

The initial probabilities associated with the existence of the various states are conveniently written as a probability vector -- a vector with as many columns as there are possible states.* For the two-parameter situation above (a_1, a_2), the probabilities of initially being in the two states are written as the vector

$$[q_1 \quad q_2] \quad ,$$

where q_1 is the probability of beginning in state a_1, and q_2 is the probability of beginning in state a_2. If an experiment always began in state a_1, then the vector would be:

$$[1.0 \quad 0] \; .$$

* A probability vector is a one-row matrix whose elements are non-negative and total 1.0.

If there were an equal chance that the experiment would start out in either of the two classes, the vector would be:

$$[.5 \quad .5] \, .$$

If this vector is multiplied by the transition matrix, it will produce the probabilities of being in the two states at the end of the first period. The product also will be a probability vector. If this second vector is multiplied by the transition matrix, it will, in turn, produce the probable states in the second period, and so on. This procedure is referred to as "chaining."

A matrix of transition probabilities (in fact any stochastic matrix), together with a set of initial probabilities, completely determines a discrete-parameter Markov chain.*

To illustrate, consider the following example: A man takes a business trip and leaves his wife at home. It is well established that wives consider business trips only as riotous interludes to a humdrum existence which, in their position of servitude, they are denied. To mollify her sense of social injustice, there is a certain probability that the wife will immediately buy a new hat. Thus, there are two possible states: h_1, in which a hat is purchased; and h_2, in which a hat is not purchased. There are probabilities associated with each, which each businessman must empirically determine for himself. Thus, the traveler anticipating a possible outcome of his trip could put these two states in a probability vector, z:

---

* A matrix is stochastic if each of the rows totals 1.0 and there are no negative terms. In this paper, only this type of matrix is discussed.

$$z = \begin{array}{cc} h_1 & h_2 \\ [.9 & .1] \end{array} .$$

In vector z, h_1 denotes "hat", and h_2 denotes "no hat."

For every week that the man is away from home, there is another set of probable events. At the end of the week, or at the beginning of the second period, there is a probability that the wife, having purchased the hat, keeps it (p_{11}); the probability that the wife, not having previously purchased the hat, does so (p_{21}); the probability that the wife purchased the hat and returned it (p_{12}); and the probability that the wife did not purchase the hat and therefore did not return it (p_{22}). (Note that the first number in the subscript pertains to the state in the first period and the second number to the state in the second period.) These probabilities form the transition matrix previously described -- i.e., the sum-total of all possible outcomes. Since the first subscript is the row subscript in the matrix notation, the rows will denote the states in the first time period k, and the columns (the second subscript) will denote the states in the second time period. For example, a transition matrix might have the following values:

$$A = \begin{array}{c} \\ h_1 \\ h_2 \end{array} \begin{array}{c} \begin{array}{cc} h_1 & h_2 \end{array} \\ \left\| \begin{array}{cc} .8 & .2 \\ .7 & .3 \end{array} \right\| \end{array} .$$

If one multiplies matrix A by the probability vector z, the result will be the probability that the wife will have a hat at the conclusion of the first time period -- i.e., one week.* Thus,

* This is row-column multiplication of matrices. For those not familiar with matrix algebra, a short, lucid discussion can be found in Mood, A. M., Introduction to the Theory of Statistics, McGraw-Hill Book Company, Inc., New York, 1950, p. 171.

$$\begin{array}{cc} h_1 & h_2 \\ [.9 & .1] \end{array} \begin{array}{cc} h_1 & h_2 \\ \end{array}\left\|\begin{array}{cc} .8 & .2 \\ .7 & .3 \end{array}\right\| = \begin{array}{cc} h_1 & h_2 \\ [.9(.8) + .1(.7) & .9(.2) + .1(.3)] \end{array}$$

$$= \begin{array}{cc} h_1 & h_2 \\ [.79 & .21] \end{array} = z^{(1)} .$$

The probability of having a hat at the end of the first week, h_1, is the probability that a hat was purchased times the probability that the hat, having been purchased, was kept (.9 x .8), plus the probability it was not purchased at the beginning of the first week times the probability it was purchased at the end of the week (.1 x .7). Thus, the probability that the wife owns a new hat at the beginning of the second week is .79.

What happens as the second week passes? If one assumes that the wife's feelings of servitude have not increased with the resulting purchase of a grand piano, an estimate of events may then be had by multiplying the second vector, $z^{(1)}$, times the transition matrix,

$$[.79 \quad .21] \left\|\begin{array}{cc} .8 & .2 \\ .7 & .3 \end{array}\right\| = \begin{array}{cc} h_1 & h_2 \\ [.779 & .221] \end{array} = z^{(2)} .$$

Thus, there is a smaller probability that the wife will own a new hat at the end of the second week than at the end of the first. (It is hoped that the logic of this outcome will not invalidate the example.) Successive multiplications of the resulting vectors times the original matrix will produce the probabilities for succeeding weeks away from home.

After a few multiplications (weeks), the businessman will note that the two values change very little between successive weeks and soon become "stabilized." This would indicate that after a while, a

longer stay will not materially change the probabilities that a hat will be purchased. For the third multiplication, $z^{(2)}A$,

$$[.779 \quad .221] \begin{Vmatrix} .8 & .2 \\ .7 & .3 \end{Vmatrix} = \begin{array}{cc} h_1 & h_2 \\ [.7779 & .2221] \end{array} = z^{(3)} .$$

These same vectors, $z^{(1)}$, $z^{(2)}$, $z^{(3)}$, ..., $z^{(n)}$, ..., can be generated by raising the matrix to successively higher powers and multiplying each power of the matrix by the original probability vector, z. Thus,

$$[.9 \quad .1] \quad A^2 = [.9 \quad .1] \begin{Vmatrix} .78 & .22 \\ .77 & .23 \end{Vmatrix}$$

$$= [.779 \quad .221] = z^{(2)}$$

$$[.9 \quad .1] \quad A^3 = [.7779 \quad .2221] = z^{(3)} ,$$

et cetera.

Also, as the stochastic matrix is raised to higher and higher powers, a_{ij} and a_{kj} approach equality for every i and k. Each row will approach the same values as the probability vector obtained by multiplying this power of the matrix by the probability vector.*

---

* The following discussion will involve only the above-mentioned matrix characteristics. A more complete discussion can be found in Kemeny, J. G., Snell, J. L., and Thompson, G. L., Introduction to Finite Mathematics, Prentice-Hall, Inc., 1958, Chapter V, Sections 7 and 8. On pages 220 and 221, there are two theorems of interest:

(1) If P is a regular stochastic matrix, then:
(a) the powers P^n approach a matrix T,
(b) each row of T is the same as the probability vector t,
(c) the components of t are positive.

(2) If P is a regular stochastic matrix and T and t are given by the previous theorem, then pP^n approaches ti whenever p is any probability vector.

This "stable state", as it will be referred to hereafter, is approached as a limit,

$$\lim_{n \to \infty} z^{(n)}A = \lim_{n \to \infty} zA^{n+1} = z^{(\infty)}.$$

However, it is not necessary to carry out this limiting process to determine $z^{(\infty)}$. It can be determined as a vector which, when multiplied on the right by the transition matrix, will reproduce itself. This means that $z^{(\infty)}$ can be found by solving a system of linear equations. If the vector $z^{(\infty)}$ is represented by $[\alpha_1 \quad \alpha_2]$, and the same stochastic matrix as in the previous example is used, we have:

$$[\alpha_1 \quad \alpha_2] \begin{Vmatrix} .8 & .2 \\ .7 & .3 \end{Vmatrix} = [\alpha_1 \quad \alpha_2] \quad .$$

By regular matrix multiplication, we obtain the system of equations,

$$.8\alpha_1 + .7\alpha_2 = \alpha_1$$
$$.2\alpha_1 + .3\alpha_2 = \alpha_2 \quad .$$

These are two linear, homogeneous equations with no unique, non-trivial solution. However, one equation can be replaced by

$$\alpha_1 + \alpha_2 = 1 \quad ,$$

which follows from the fact that $z^{(\infty)}$ is a probability vector. This gives the system

$$.2\alpha_1 - .7\alpha_2 = 0$$
$$\alpha_1 + \alpha_2 = 1 \quad .$$

The solution of these two linear equations yields $\alpha_2 = 2/9$ and $\alpha_1 = 7/9$, or $\alpha_2 = .2222 \ldots$ and $\alpha_1 = .7777 \ldots$. Note that these values are not greatly different from those in $z^{(2)}$ and $z^{(3)}$. This fact indicates very rapid convergence in this Markov chain.

2.1.1 Markov Process and Continuous Model

This Markov process is clearly a discrete one, consisting of finite steps or trials. Such a model is ideal for describing some phenomena which occur in distinct steps rather than continuously with time. Experiments in genetics often furnish good examples. A given mating will produce an offspring with probabilities of certain given characteristics. The mating of the offspring -- a finite step -- will produce the given characteristics with another set of probabilities. In the field of electronics, the return of a signal on each rotation of a radar antenna -- the blip/scan ratio -- has been described by a Markov process. Here each turn of the radar forms a distinct "trial." However, in the equipment deterioration phenomenon considered here, the situation is no longer discrete but is a continuous function of time and the values which appear in the transition matrix are dependent upon the time interval selected. Consider deterioration from state a_1 to state a_2. The value of p_{12} would be considerably smaller if determined over an interval of one hour, than if determined over an interval, say, of one month. In using a Markov process in deterioration models one must be aware of this discrete aspect. This alone should not present great difficulties, for continuous phenomena have long been approximated by discrete methods and it is a natural consequence of periodic rather than continuous monitoring of equipment performance. But it means that the selection of the time interval for developing the transition matrix must be made with some consideration of this inherent discrete characteristic.

Attention is called to another aspect of the Markov process. The development of each step in the Markov chain uses only the information (the probable states) which existed in the immediately

preceding time period and no other. In other words, the states existing prior to the immediately preceding one are not drawn upon for information, nor is the manner in which the preceding state was reached used as contributing information. This is stated by Feller (p.337) -- "Conceptually, a Markov process is the probabilistic analogue of the process of classical mechanics where the future development is completely determined by the present state and is independent of the way in which the present state has developed."

If the deterioration process preceding a measurement has been exceedingly rapid, or conversely slow, the incorporation of this information should contribute to a better prognostication of system performance. To the degree that this information is not used, the first-order Markov process may leave room for improvement in actual application.* Perhaps similar application of higher order processes may correct this, but this is not believed to be a critical deficiency in the model in view of its intended application. The procedure to be used here does make use of a significant portion of the available information and later ARINC Research Corporation studies will examine methods by which the additional information might be incorporated.

2.1.2 Application of the Model to a Theoretical Problem

The transition matrix is determined experimentally in the manner described in the appendix. The probabilities p_{ij} are determined from periodic measurements of the performance of a number of equipments.

---

* The Markov process described in the previous example is a first-order or simple Markov process. Higher order Markov processes are those in which the transition probabilities depend on two or more preceding time periods. See Doob, J. L., Stochastic Processes, p.89, John Wiley, 1953.

At each measurement period, the number of equipments which have moved from one state to another is recorded. The number of levels of equipment-states is a matter of instrumentation and judgment. More sensitive instruments will permit the selection of more classes or states into which the variable can be divided. Generally, the more classes there are in the transition matrix the better, since we are approximating continuous phenomena by discrete steps. However, a large number of classes may require that processing be done by machine, since the matrix computations are certain to be laborious. For purposes of explanation, assume only four classes or states of system performance. Let state a_1 be peak performance, state a_2 be intermediate performance, state a_3 be marginal performance, and state a_4 be failure. A series of observations of equipment performance would yield the following matrix:

$$A = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{array} \begin{array}{c} \begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \end{array} \\ \left\| \begin{array}{cccc} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{array} \right\| \end{array} .$$

The interval selected for collection of data must be sufficiently short that only a few equipments will "skip" classes as their performance deteriorates. What this interval is will depend upon experimentation and experience with the equipment. The error resulting from selection of too short an interval will not be large. However, if too long an interval is selected, "second generation" equipments which have been repaired and returned to higher levels of operation may materially affect the data. This problem is covered in the appendix.

2.1.3 Classification of Failures in the Transition Matrix

In the preceding discussion, operating state a_4 was defined as failure. However, equipments which actually have values higher than a_4 may be removed from service if operating personnel are dissatisfied with performance. Thus, when the transition matrix is prepared from experimental data, class a_4 will have a dual meaning. For example, if an equipment has measured performance in class a_2, but is known to have been ordered removed from service by the operator during the interval, this equipment is classed in state a_4. Unless there is perfect functional dependence between the measured characteristic and the frequency of failure, state a_4 will include a combination of equipments -- those actually observed to be in state a_4 at the end of the interval and all other equipments removed for repair, regardless of their measured level of operation.

This failure classification can be described another way. Suppose that the transition matrix is based solely on the measured values of equipment performance, and the same four levels of the characteristic are assumed. The matrix can be written as

$$A = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{array} \begin{array}{c} \begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \end{array} \\ \left\| \begin{array}{cccc} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ 0 & 0 & 0 & 1 \end{array} \right\| \end{array} .$$

With each level of performance, there is a certain probability of failure, and as the performance level decreases, we expect an increasing probability of failure. Thus there are failure probabilities associated with each state:

$$\begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \\ [\, f_1 & f_2 & f_3 & 1\,] \end{array} .$$

Again it is assumed that the last measured state unmistakably represents an equipment failure.

The probability that an equipment in state a_1 is not declared a failure is $1 - f_1$. A failure/non-failure matrix can be written:

$$F = \begin{Vmatrix} 1 - f_1 & 0 & 0 & f_1 \\ 0 & 1 - f_2 & 0 & f_2 \\ 0 & 0 & 1 - f_3 & f_3 \\ 0 & 0 & 0 & 1 \end{Vmatrix} .$$

Then,

$$AF = \begin{Vmatrix} p_{11}(1-f_1) & p_{12}(1-f_2) & p_{13}(1-f_3) & p_{11}f_1+p_{12}f_2+p_{13}f_3+p_{14} \\ p_{21}(1-f_1) & p_{22}(1-f_2) & p_{23}(1-f_3) & p_{21}f_1+p_{22}f_2+p_{23}f_3+p_{24} \\ p_{31}(1-f_1) & p_{32}(1-f_2) & p_{33}(1-f_3) & p_{31}f_1+p_{32}f_2+p_{33}f_3+p_{34} \\ 0 & 0 & 0 & 1 \end{Vmatrix} .$$

In this example, $p_{11}(1 - f_1)$ is the probability that after one time interval the equipment will remain in state a_1 if failure has not occurred; $p_{12}(1 - f_2)$ is the probability that after one time interval the equipment will have moved to state a_2, again if failure has not occurred; and so on. The fourth column gives the failure probabilities in the dual sense. This type of transition matrix is the one which is dealt with in the remainder of the paper.

2.1.4 The Extension to More Than One Performance Characteristic

The failure probability vector has two extremes, which for four measured levels are

$$[\,0 \quad 0 \quad 0 \quad 1\,]$$

and

$$[\,1/4 \quad 1/4 \quad 1/4 \quad 1/4\,]$$

In the first extreme, failure is functionally dependent upon the measured performance characteristic; in the second, when each entry is 1/4, the measured performance characteristic is of no value in the prediction of failure and another characteristic should be sought. Between these two extremes, a combination of characteristics is suggested -- that is, a second measured characteristic may account for causes of failure not covered by the first one. It is possible to form a system of states based upon a combination of measurements of the two characteristics in the following manner. Suppose, for example, that the states of the initial characteristic are a_1, a_2, a_3, and a_4, and that the states of the second characteristic are b_1, b_2, and b_3. This array of combined states may be written as

$$\begin{array}{c c} & \begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \end{array} \\ \begin{array}{c} b_1 \\ b_2 \\ b_3 \end{array} & \left\| \begin{array}{cccc} c_1 & c_2 & c_3 & c_4 \\ c_5 & c_6 & c_7 & c_8 \\ c_9 & c_{10} & c_{11} & c_{12} \end{array} \right\| \end{array}$$

Thus, c_7 represents state a_3 for the first characteristic and state b_2 for the second characteristic. It is now possible to treat the c_i as a new variable with 12 states, by use of the methods described herein. The extension to more than two characteristics is obvious.

2.2 Numerical Example of the Model

Consider a numerical example of a transition matrix.*

---

* This matrix was developed as the matrix AF in Section 2.1.3. For convenience it will be denoted simply by A in the remainder of the paper.

$$A = \begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \\ \end{array}$$

$$A = \begin{Vmatrix} .5 & .5 & 0 & 0 \\ 0 & .5 & .5 & 0 \\ 0 & 0 & .1 & .9 \\ 0 & 0 & 0 & 1.0 \end{Vmatrix} .$$

This matrix has the property that once an equipment progresses into a lower state it will not, in the interval, go to a higher state, i.e., repair itself. Thus A is a triangular matrix in which the p_{ij} are all zero for $i > j$. (Note also that $p_{44} = 1.0$. If a set begins a period in the failure state, a_4, it will be there at the end of period.)

Usually, one would expect a transition matrix to have non-zero values below the main diagonal. However, if the characteristic being measured is really a deterioration phenomenon, the probabilities of transition to higher performance states should be quite small. The matrix should tend to be triangular in the sense that the values of p_{ij} should be very small if $i > j$. Furthermore, one would usually expect p_{ij} to be quite small if i is much smaller than j. In summary, this means that the performance characteristic selected for measurement should be one for which improvement is rare, deterioration is common, and states are so defined that equipments do not commonly deteriorate more than one state in the basic time interval of the transition matrix. All three of these properties are satisfied by the numerical example selected. The first condition is satisfied since p_{21}, p_{31}, p_{32}, p_{41}, p_{42}, and p_{43} are all zero. The non-zero values of p_{11}, p_{12}, p_{22}, p_{23}, p_{33}, p_{34}, and p_{44} are consistent with the second property. Finally, the zero values for p_{13}, p_{14}, and p_{24}

reflect the third property. It should be stressed, however, that the method can be used to develop state distributions no matter what the form of the stochastic transition matrix.

From an engineering viewpoint, it is important to realize that the assumptions of deterioration and accurate instrumentation naturally lead toward a triangular matrix. If the experimental data do not reflect this, it would suggest that the measured characteristic is not a good one on which to base prediction of failure, or that the accuracy of measurement is too crude to monitor equipment performance, or that both of these conditions hold. In this case, the situation must be examined to see if another characteristic must be selected, or if instrumentation can be improved.

We can, from this information, generate the failure density function of the equipment. If the equipments all begin in state a_1, then the initial probability vector is:

$$\begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \\ [\,1 & 0 & 0 & 0\,] \end{array} .$$

If this vector is multiplied by the transition matrix, the distribution by state at the end of one time interval is obtained. Repetition of this process generates the state distribution over time as shown in Table 1. Columns 2 through 5 give the state distribution for the times shown in Column 1. This constitutes an important description of the equipment deterioration pattern, based entirely on the transition matrix. The failures, which were identified as state a_4, are shown in Column 5 in the form of the unreliability function, $U(t)$. This results from the fact that the process includes no repair, so the cumulative failure frequency is developed. The failure density

TABLE 1

COMPUTATIONS FOR THE NUMERICAL EXAMPLE

Time (t)	State a_1	a_2	a_3	a_4 Failure U(t)	Failure-Density Function u(t)	Reliability Function R(t)	Average State Values
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0	1	0	0	0		1	1.00
1	.5	.5	0	0	0	1	1.50
2	.25	.5	.25	0	0	1	2.00
3	.125	.375	.275	.225	.225	.775	2.60
4	.0625	.250	.215	.4725	.2475	.5275	3.10
5	.03125	.15626	.1465	.6660	.1935	.3340	3.45
6	.015625	.09375	.092775	.79785	.13185	.20215	3.67
7	.0078125	.0546875	.0561525	.8813475	.0384675	.1186425	3.81
8	.00390625	.03125	.032959	.93188475	.05053725	.06811525	3.89
9	.001953125	.017578125	.0189209	.96154785	.02966310	.03845215	3.94
10	.0009765625	.009765625	.0106811525	.97857666	.01702881	.02142334	3.97
11	.00048828125	.00537109375	.00595092775	.98818969725	.00961303725	.00181030275	3.98
t > 11	0	0	0	1.000	.01181030275	0	4.00

function, $u(t)$, in Column 6, gives the probability of failure in each time period. (See Figure 2.) It is obtained by taking differences between $U(t)$ values of Column 5. The reliability function, shown in Column 7 is computed from Column 5 by the formula

$$R(t) = 1 - U(t) \ .$$

The average level of measured equipment performance was basic in the statement of the problem given earlier in this report. Such an average level is in reality an "average state" as a function of time. In order to compute such an average state and to draw a graph for the transition matrix, it is necessary to identify states numerically instead of just by names a_1, a_2, ..., a_5. Column 8 shows average state based on the assignment of numerical value 1 to a_1, 2 to a_2, etc. Thus, each entry in Column 8 is the average state of the probability vector in Columns 2 through 5 for the time shown in Column 1.

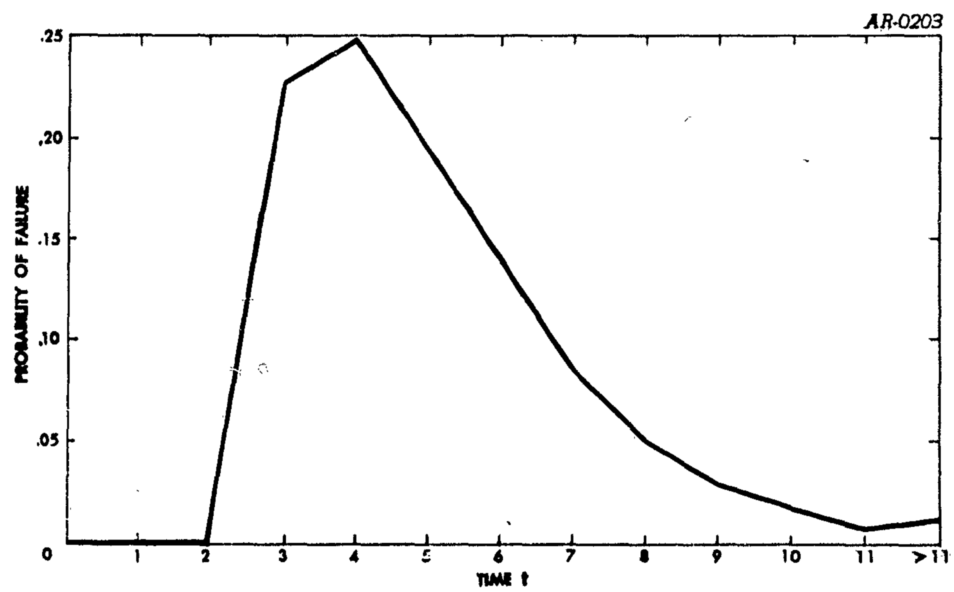

FIGURE 2
TIME-TO-FAILURE DENSITY FUNCTION, u(t)

Figure 3 is a graph of the average state values in Column 8 plotted against time, Column 1. In addition, the state distribution for time t = 3 is shown in the upper right hand corner of the figure.

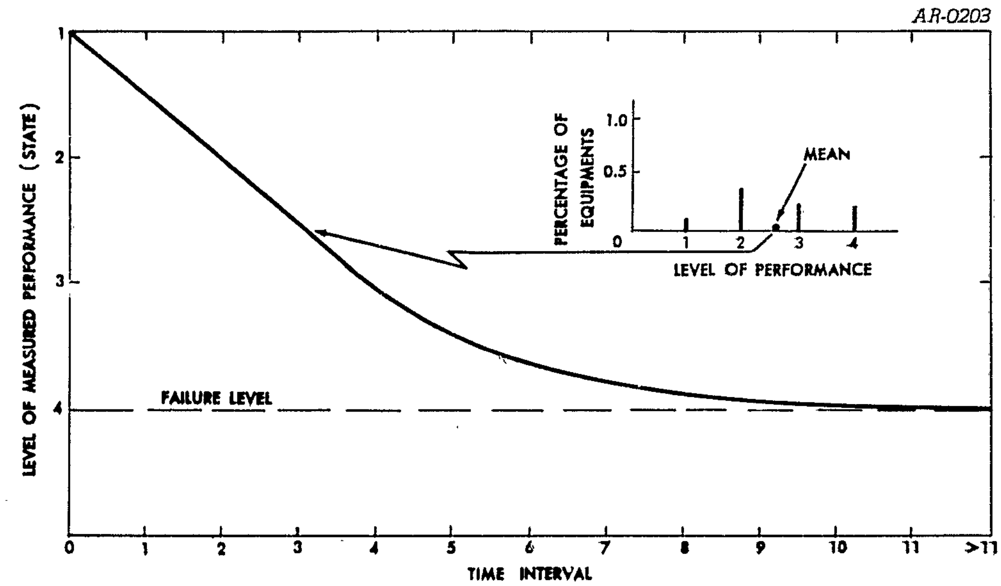

FIGURE 3
AVERAGE DETERIORATION OF A PERFORMANCE CHARACTERISTIC

It should be noted that the method of selecting a preventive maintenance criterion is not dependent on the assignment of numerical values to the various states. Numerical values are assigned* here only in the desire to describe the deterioration phenomenon by means of the transition matrix, and to relate the computations in the remaining portion of this report to the problem stated in terms of deterioration.

---

* In most cases, the numerical values would be given directly by the characteristic measurement.

2.3 Scheduling of Preventive Maintenance

The foregoing discussion is concerned primarily with the development of the basic concept that deterioration phenomena can be represented by a Markov process. Reference has been made to the problems associated with data collection, definition of suitable performance levels, and determination of a time interval for the fundamental transition matrix consistent with repair-time requirements.

At this point, it is necessary to describe methods whereby matrices can be modified to provide for repair of in-service failure at times consistent with normal maintenance practices, and also to provide for independently scheduled preventive maintenance. It is reasonable to assume that preventive maintenance will be scheduled at intervals which are long compared to the time required for the repair of an in-service failure.

When the unit of time is the interval covered by the basic transition matrix, the problem is to develop a method for computation of the expected number of in-service failures which will occur in n time unit intervals -- with repair of in-service failures at the end of each unit interval and preventive maintenance at the end of the n^{th} unit interval -- and to express the entire process in matrix form. It is assumed here that maintenance always occurs at the end of each unit interval.

For the sake of convenience, the numerical matrix shown on page 20 is repeated here. It will be recalled that state a_4 constitutes failure.

$$A = \begin{array}{c} \begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \end{array} \\ \left\| \begin{array}{cccc} .5 & .5 & 0 & 0 \\ 0 & .5 & .5 & 0 \\ 0 & 0 & .1 & .9 \\ 0 & 0 & 0 & 1.0 \end{array} \right\| \end{array} .$$

If it is assumed that repair of an in-service failure returns the equipment to the highest operating state, the basic matrix is modified by adding Column 4 to Column 1 to form a new Column 1. Since such repairs are made at the end of each time interval, the basic transition matrix becomes:

$$A_{-1} = \left\| \begin{array}{ccc} .5 & .5 & 0 \\ 0 & .5 & .5 \\ .9 & 0 & .1 \end{array} \right\| .$$

Note that Column 4 and Row 4 of Matrix A have been deleted. This is done here because the repair of in-service failures precludes the existence of equipments in state a_4, and therefore a matrix with three rows and three columns is adequate to describe the situation.*

---

* It is of interest to note that this operation can be expressed as a matrix product. This leaves a four by four matrix containing A_{-1} and indicating the shift of equipments from a_4 to a_1.

$$\left\| \begin{array}{cccc} .5 & .5 & 0 & 0 \\ 0 & .5 & .5 & 0 \\ 0 & 0 & .1 & .9 \\ 0 & 0 & 0 & 1.0 \end{array} \right\| \left\| \begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{array} \right\| = \left\| \begin{array}{cccc} .5 & .5 & 0 & 0 \\ 0 & .5 & .5 & 0 \\ .9 & 0 & .1 & 0 \\ 1.0 & 0 & 0 & 0 \end{array} \right\| .$$

Thus, A is multiplied on the right by a matrix expressing that the probability is 1.0, that state a_1 will remain a_1, that a_2 will remain a_2, that a_3 will remain a_3, and a_4 will become a_1.

Deletion of one row and one column is indicated by the subscript in the symbol A_{-1}. The notation will be extended later as the matrix is reduced by deletion of additional rows and columns.

If the process of repairing in-service failures is continued indefinitely, a stable state is approached,* as illustrated at the end of Section 2.1. If the stable state is

$$z^{(\infty)} \equiv [\alpha_1 \ \ \alpha_2 \ \ \alpha_3] \ ,$$

then

$$z^{(\infty)} A_{-1} = z^{(\infty)} \ .$$

This gives the system of equations

$$\begin{aligned} .5\,\alpha_1 + .9\,\alpha_3 &= \alpha_1 \\ .5\,\alpha_1 + .5\,\alpha_2 &= \alpha_2 \\ .5\,\alpha_2 + .1\,\alpha_3 &= \alpha_3 \\ \\ \alpha_1 + \alpha_2 + \alpha_3 &= 1. \end{aligned}$$

Therefore, $z^{(\infty)} = [.3913 \quad .3913 \quad .2174]$.

In the stable-state condition, the expected number of in-service failures in one time interval can be obtained by multiplying the vector $z^{(\infty)}$ by the fourth column of matrix A:

$$z^{(\infty)} A = \begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \\ [.3913 & .3913 & .2174 & 0] \end{array} \left\| \begin{array}{ccc|c} .5 & .5 & 0 & 0 \\ 0 & .5 & .5 & 0 \\ 0 & 0 & .1 & .9 \\ 0 & 0 & 0 & 1.0 \end{array} \right\| .$$

The fourth element in the product is $(.9)(.2174) = .1957$. Therefore, this is the expected frequency of in-service failures in each time interval after the stable state has been reached.

---

* If no repairs are made, the stable state developed from matrix A is [0 0 0 1].

2.3.1 Computations for a Maintenance Cycle of n Time Intervals

To illustrate the computational procedure, suppose preventive maintenance were performed at the end of every sixth interval. Then the total number of maintenance actions would be the sum of repairs of in-service failures -- performed at the end of each interval -- and the required preventive maintenance actions at the end of the sixth interval. With four states, a_1, a_2, a_3, and a_4, with a_4 being failure, the a_4 would be shifted by repair of in-service failure to a_1 at the end of each interval. Preventive maintenance would consist of transfer of a_3 (or a_3 and a_2) to a_1 at the end of every sixth time interval.

These maintenance actions can be expressed in matrix notation in the manner shown below.

Let

$$x \equiv [x_1, x_2, x_3]$$

denote the probability vector for the initial distribution of equipments by states. The transition matrix A_{-1} describes the state transition in one time interval if in-service failures are repaired but no preventive maintenance is performed.* The probability vector at the end of one time interval is the product xA_{-1}. This vector becomes xA_{-1}^n at the end of the n^{th} interval. In applying this method it is convenient to compute powers of A_{-1}. The powers of interest in this example are shown below.

---

* As indicated in Section 2.3, the notation A_{-1} was adopted to denote the deletion of one row and one column following the addition of Column 4 to Column 1. If the preventive maintenance schedule called for repair of equipments in state a_3, two columns would be added to Column 1, and the notation would be A_{-2}; etc.

$$A_{-1} = \begin{Vmatrix} .5 & .5 & 0 \\ 0 & .5 & .5 \\ .9 & 0 & .1 \end{Vmatrix}$$

$$A^2_{-1} = \begin{Vmatrix} .25 & .50 & .25 \\ .45 & .25 & .30 \\ .54 & .45 & .01 \end{Vmatrix} \qquad A^3_{-1} = \begin{Vmatrix} .350 & .375 & .275 \\ .495 & .350 & .155 \\ .279 & .4[illegible]5 & .226 \end{Vmatrix}$$

$$A^4_{-1} = \begin{Vmatrix} .4225 & .3625 & .2150 \\ .3870 & .4225 & .1905 \\ .3429 & .3870 & .2701 \end{Vmatrix} \qquad A^5_{-1} = \begin{Vmatrix} .40475 & .39250 & .20275 \\ .36495 & .40475 & .23030 \\ .41454 & .36495 & .22051 \end{Vmatrix}$$

$$A^6_{-1} = \begin{Vmatrix} .38485 & .39862 & .21652 \\ .38974 & .38485 & .22540 \\ .40572 & .38974 & .20452 \end{Vmatrix} \qquad A^7_{-1} = \begin{Vmatrix} .38729 & .39175 & .22096 \\ .39773 & .38730 & .21496 \\ .38694 & .39773 & .21538 \end{Vmatrix}$$

$$A^8_{-1} = \begin{Vmatrix} .39252 & .38952 & .21796 \\ .39234 & .39252 & .21514 \\ .38726 & .39234 & .22040 \end{Vmatrix} .$$

It will be noted that the rows converge toward the stable-state vector

$$s^{(\infty)} = [\,.3913 \quad .3913 \quad .2174\,] \quad ,$$

which was derived by the method described in Sections 2.1 and 2.3.

If the preventive maintenance schedule requires the transfer of state a_3 equipments to state a_1 at the end of n intervals, the stable-state vector at the end of the n^{th} interval is determined by the same methods. For example, if n = 4 (note A^4_{-1}), the transition matrix is

$$A_{-2} = \begin{Vmatrix} .4225 + .2150 & .3625 \\ .3870 + .1905 & .4225 \end{Vmatrix} = \begin{Vmatrix} .6375 & .3625 \\ .5775 & .4225 \end{Vmatrix} ,$$

which, when multiplied on the left by the stable-state probability vector $[\alpha_1 \; \alpha_2]$, reproduces this vector. Thus

$$[\alpha_1 \quad \alpha_2] \begin{Vmatrix} .6375 & .3625 \\ .5775 & .4225 \end{Vmatrix} = [\alpha_1 \quad \alpha_2] .$$

This is equivalent to

$$.6375\,\alpha_1 + .5775\,\alpha_2 = \alpha_1$$
$$.3625\,\alpha_1 + .4225\,\alpha_2 = \alpha_2$$

Replacing one equation by $\alpha_1 + \alpha_2 = 1$, the solution is found to be

$$\alpha_1 = .614, \alpha_2 = .386.$$

Therefore, the stable-state vector in this instance is

$$[.614 \quad .386] .$$

This is really an abbreviation for the vector [.614 .386 0 0], which indicates that no equipments are left in states a_3 and a_4 from one preventive-maintenance interval to another. It must be remembered, however, that the equipments do pass through these states in the intervals between preventive maintenance actions.

Stable-state probability vectors for a selection of preventive maintenance schedules are shown below:

Preventive-Maintenance Interval (n)	Stable-State Vector			
n	a_1	a_2	a_3	a_3
1	[.5	.5	0	0]
2	[.6	.4	0	0]
3	[.634	.366	0	0]
4	[.614	.386	0	0]
. . . .	. . . .	. . . .	. . .	. . .
8	[.6381	.3619	0	0].

The number of failures which occur in the time interval between preventive maintenance actions is determined by chaining the n^{th} stable-state vector, using multiplication by matrix A "n" times.* This procedure is illustrated for the case where preventive maintenance occurs every fourth interval.

Period										
1	[.614	.386	0	0]A =	[.3070	.5000	.1930	0]	=	$z^{(1)}$
2	[.3070	.5000	.1930	0]A =	[.15350	.40350	.2693	.1737]	=	$z^{(2)}$
			.1737 + .15350 = .3272**							
3	[.3272	.4035	.2693	0]A =	[.16360	.36535	.33868	.24237]	=	$z^{(3)}$
4	[.40597	.36535	.22868	0]A =	[.20299	.38566	.20554	.20581]	=	$z^{(4)}$.

The total number of in-service failures which occur during each four-interval period between preventive maintenance actions is 0 + .1737 + .2434 + .2058 = .6229. The number of equipments which undergo preventive maintenance, those in state a_3 at the end of the fourth period, is .20554.

The total number of in-service failures and preventive maintenance replacements for other groups of intervals are determined in a similar fashion. The chaining occurs a different number of times and the associated stable states are different for different intervals. Computations for preventive maintenance every sixth interval are given below.

---

* See procedure given on page 27.

** During the second interval, .1737 failures occurred. Since it was hypothesized that these equipments were repaired during the interval (restored to state a_1), there is no state a_4 beginning with the third interval.

$$
\begin{aligned}
z^{(0)} &= [\,.6067 \quad .3932 \quad 0 \quad 0\,] \\
z^{(1)} &= [\,.3034 \quad .5000 \quad .1966 \quad 0\,] \\
z^{(2)} &= [\,.3286 \quad .4017 \quad .2697 \quad (.1769)^{*}\,] \\
z^{(3)} &= [\,.4070 \quad .3652 \quad .2278 \quad (.2427)\,] \\
z^{(4)} &= [\,.4085 \quad .3861 \quad .2054 \quad (.2050)\,] \\
z^{(5)} &= [\,.3891 \quad .3973 \quad .2136 \quad (.1849)\,] \\
z^{(6)} &= [\,.3868 \quad .3932 \quad .2200 \quad (.1922)\,]
\end{aligned}
$$

n_1 = 1.0017 = number of in-service failures

n_2 = .2200 = number of preventive replacements.

Table 2 gives the number of in-service failures (n_1) and the number of equipments replaced during preventive maintenance (n_2) for various values of n, the number of intervals between preventive maintenance actions.

TABLE 2

NUMBER OF IN-SERVICE FAILURES AND PREVENTIVE REPLACEMENTS, WHEN PREVENTIVE MAINTENANCE IS SCHEDULED EVERY SIXTH INTERVAL

Interval (n)	In-Service Failures (n_1)	Preventive Replacements (n_2)
1	0	.25
2	.18	.27
3	.4062	.2311
4	.6219	.2055
5	.8109	.2137
6	1.002	.2200
7	1.3842	.2169

2.3.2 Summary of the Mathematical Model

The mathematical model developed in preceding sections expresses the deterioration pattern of an equipment in the form of a Markov

* State a_4 is parenthesized because these values have been added to state a_1.

process. This transition matrix covers an interval of time which is sufficiently short to preclude the influence of second-generation equipments -- that is, equipments which are repaired and returned to service before the end of the interval.* The distribution of equipments by performance level or state at any subsequent time is expressed as the product of the initial state distribution and an appropriate power of the transition matrix. This power is equal to time, expressed in units of the interval for which the transition matrix is applicable.

Maintenance procedures can be expressed as modifications of the transition matrix. Repair of in-service failures is reflected by the addition of the failure-state column to the column representing the state following repair. Preventive maintenance is reflected by the addition of lower-state columns to higher-state columns as appropriate. In the present discussion, it is always assumed that preventive maintenance and repair of in-service failures restore the equipment to the highest state, which is represented by the first column of the matrix.

If in-service failures are repaired immediately after occurrence (it is implicitly assumed that they will be repaired within the time period covered by the transition matrix) and equipments in certain states lower than state a_1 are repaired at the end of every n^{th} interval, a stable state is developed around this replacement pattern.

---

* The selection of appropriate intervals is discussed in Section 1 of the Appendix.

Each replacement pattern will generate a different number of failures since each will have its own stable state. The arithmetic is expressed in terms of matrix multiplication. By computing several replacement patterns which differ both as to time interval and replacement level, it is possible to compare the total number of in-service failures and preventive replacements generated by each maintenance pattern. This comparison provides the data required to determine which pattern of maintenance yields the lowest cost. All cost computations can be based on stable-state distributions, since the ultimate average cost is independent of the initial state distribution.

3. DETERMINATION OF OPTIMUM MAINTENANCE SCHEDULES

3.1 The Cost Equation

The expected average cost per unit time is given by the equation:

$$C = \frac{1}{h} \left[k_1 n_1 + k_2 n_2 + k_3 \right] .$$

where:

h = the number of time units between periodic preventive maintenance actions; one time unit is the period of time for which the transition matrix is developed.

k_1 = the cost of repair of an in-service failure.

k_2 = the cost of a scheduled preventive maintenance action.

k_3 = the cost of periodic test or measurement of performance level.

n_1 = the expected number of in-service failures in h time units.

n_2 = the expected number of scheduled preventive maintenance actions in h time units.

The situations of interest are those in which the cost of repairing an in-service failure is considerably greater than the cost of a preventive maintenance action at scheduled maintenance intervals (there would be little reason for preventive maintenance if it were more expensive than repair of an in-service failure). Therefore, in the numerical illustrations which follow, it is assumed that the values of k_1 are considerably larger than the values of k_2. It is also assumed that the values of k_3 (the cost of making the periodic check of equipment performance) are less than either of the other costs, although this is not a necessary assumption and has no effect on the validity of the method.

3.2 A Four-State Example

To illustrate the method, cost computations for the above numerical example are made for a selection of cost parameters. Assume two sets of values:

$$(1)\quad k_3 = 1, \quad k_2 = 4, \quad k_1 = 8$$
$$(2)\quad k_3 = 1, \quad k_2 = 4, \quad k_1 = 16.$$

If no preventive maintenance is performed -- that is, if equipment is repaired only after failure -- there will be a constant failure rate of .1957 per time interval (see page 27), and the cost equation is

TABLE 3

AVERAGE COST PER UNIT TIME WHEN PREVENTIVE MAINTENANCE IS PERFORMED EVERY n^{th} INTERVAL BY REPLACING EQUIPMENTS IN STATE a_3

Replacement Interval	Cost Equation	Cost Schedule 1* $k_1 = 8$	Cost Schedule 2* $k_1 = 16$
1	$\frac{1}{1}\ [0k_1 + .25k_2 + k_3]$	2.00	2.00
2	$\frac{1}{2}\ [.18k_1 + .27k_2 + k_3]$	1.76	2.48
3	$\frac{1}{3}\ [.4062k_1 + .2311k_2 + k_3]$	1.72	2.81
4	$\frac{1}{4}\ [.6219k_1 + .2055k_2 + k_3]$	1.70	2.94
5	$\frac{1}{5}\ [.8110k_1 + .2136k_2 + k_3]$	1.66	2.97
6	$\frac{1}{6}\ [1.002k_1 + .2200k_2 + k_3]$	1.65	2.98
8	$\frac{1}{8}\ [1.384k_1 + .2169k_2 + k_3]$	1.62	3.00

* In both cost schedules, $k_2 = 4$ and $k_3 = 1$.

$$C = \frac{1}{1} \; [.1957\, k_1 + 0k_2 + 0k_3] = 1.56, \text{ if } k_1 = 8$$
$$= 3.13, \text{ if } k_1 = 16.$$

Assume preventive maintenance is performed every n^{th} interval by restoring equipments in state a_3 to state a_1. Then the resulting cost equations and total costs are those shown in Table 3. The costs listed in the table are plotted in Figure 4. These curves lead to the following observations. When Schedule 1 costs are assumed, preventive maintenance at any time interval is more costly than none at all (the cost of repair of in-service failures with no preventive maintenance is indicated by the horizontal line marked $k_1 = 8$). On the other hand, when Schedule 2 costs are assumed, any preventive maintenance is better than none, irrespective of time -- the optimum situation being obtained when preventive maintenance is performed at the end of the first measurement interval.

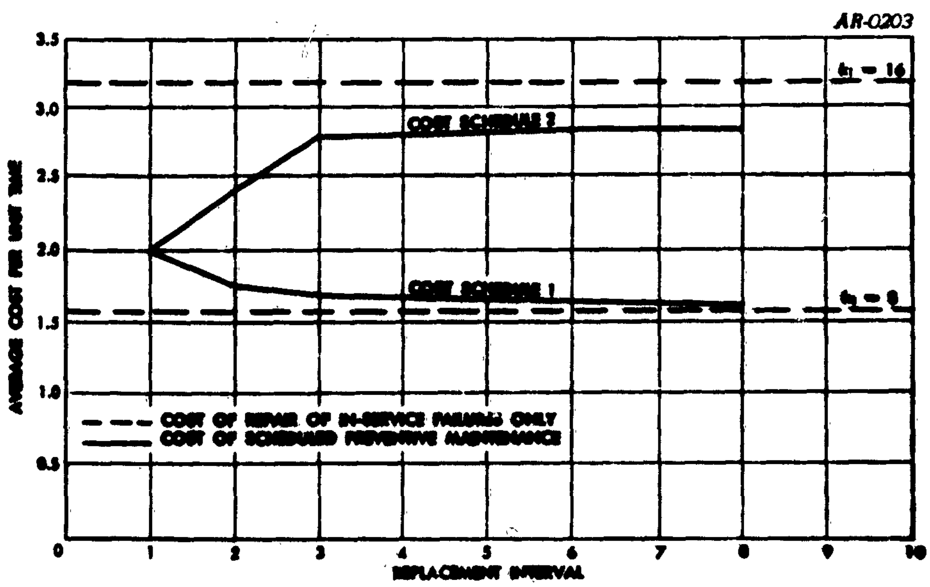

FIGURE 4
COST OF REPAIR OF IN-SERVICE FAILURE vs.
COST OF SCHEDULED PREVENTIVE MAINTENANCE

By a similar procedure, the relative costs of other patterns of preventive maintenance -- for example, replacement of equipments in states a_3 and a_2 -- can be determined. Such a comparison would be necessary for a complete cost analysis. However, in view of the small number of performance levels involved here, further study of this simple example is unwarranted.

Intuitively, one would expect the optimum time for replacement to occur somewhere between the first and the infinite intervals; i.e., there would seem to be a few actual situations in which maintenance "as early as possible" or "not at all" would be warranted. In fact, the small simple matrix used in the preceding example would fit few actual situations. Because of the values selected and the small number of classes used, a stable state is reached very quickly, which forces second- and third-generation failures to enter rapidly into the average failure rate. Table 2 (see page 32) indicates that failures begin in the second interval, that nearly two-thirds of them have occurred by the end of the fourth interval, and that new-generation failures have occurred by the end of the sixth interval. Thus, the equipments rapidly reach a random-age distribution, and the failure density function for the second generation so overlaps the first-generation density that the cost curves (which reflect system failures) are quite smooth.*

---

* Welker, Dr. E. L., Relationship Between Equipment Reliability, Preventive Maintenance Policy, and Operating Costs, ARINC Research Corporation, February 13, 1959 (Publication No. 101-9-135), pp. 20 ff.

t should also be noted that the curves shown in Figure 4 begin at the first time interval. All cost curves start with this interval, as cost equations and number of failures can be determined only for intervals which are at least as long as the one selected for the transition matrix. If the basic time interval in this example were shortened, the beginning value for both curves in the figure would be considerably higher, because the cost of nearly continuous checking (k_3) would be much higher over any given interval of time.

3.3 A Seven-State Example

To illustrate a more typical case, another example is given. In this example, the underlying density function has a smaller coefficient of variation, a property which will turn out to be critical in developing a more usual deterioration pattern.

$$B = \begin{array}{c} \begin{array}{ccccccc} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 \end{array} \\ \begin{Vmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & .5 & .5 & 0 & 0 \\ 0 & 0 & 0 & 0 & .5 & .5 & 0 \\ 0 & 0 & 0 & 0 & 0 & .1 & .9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1.0 \end{Vmatrix} \end{array} .$$

It will be noted that the submatrix shown in the lower right-hand corner is identical to the one used in previous illustrations. Appending the first three rows and columns has the effect of shifting the failure density to the right by three time intervals (see Figure 3). Therefore, the mean of the distribution is now 8.1 time units, but the standard deviation of 2.0 is unchanged from the original

example.* The net effect of this shift is to decrease the coefficient of variation from .40 to .25. The additional columns can be regarded as additional states. State a_7 now designates failures which, when repaired, are returned to state a_1.

3.3.1 "Dummy" Columns in the Transition Matrix

In actual practice, it may be necessary to add "dummy" columns to the transition matrix derived from empirical data. Whether or not this is done depends upon the testing instruments used in the experiment. For example, if these are sufficiently sensitive to measure seven rather than four classes of performance, there may be values other than zero or one in the first three columns. On the other hand, if the equipment is a receiver with a considerable number of redundant elements, it will probably register performance close to peak levels— state a_1 -- for a long period of time and then fail quite rapidly. In this situation, the average operating level would remain almost constant and then drop off sharply, as shown in Figure 5.

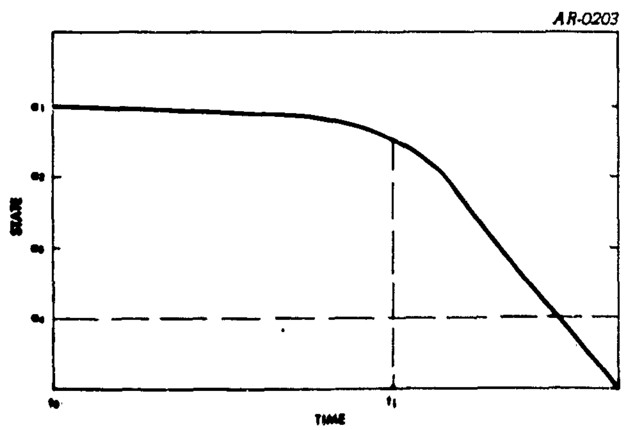

FIGURE 5
AVERAGE OPERATING LEVEL AS A FUNCTION OF TIME

* See Appendix, Section 4

Actually, under the hypothesis that failure is a gradual deterioration phenomenon, it can be assumed that the deterioration shown in Figure 5 began at time t_0 but that the instrument was unable to measure it. Thus, dummy columns would be required to shift the failure density function to the right, as in the case of matrix B. The portion of the curve to the right of t_1 is analogous to the entire curve shown in Figure 2. This matter is discussed in more detail in Section 2 of the Appendix.

3.3.2 Cost Equations for Matrix B

The computations associated with matrix B are carried out in the same manner as those of the previous example. Addition of Column 7 to Column 1 constitutes repair of in-service failures, while addition of any other columns to Column 1 constitutes preventive maintenance. Cost equations for preventive maintenance at intervals up to 10 are shown in Tables 4 and 5 -- Group I equations for maintenance involving replacement of equipments in state a_6, and Group II equations for maintenance involving replacement of equipments in states a_5 and a_6.

The following sets of costs are used in the tables.

Cost	Schedule 1	Schedule 2	Schedule 3	Schedule 4
k_1	8	11	15	19
k_2	4	4	4	4
k_3	1	1	1	1

The cost equation and maintenance costs for repair of in-service failures only (no preventive maintenance actions) are:

TABLE 4

GROUP I COST EQUATIONS: AVERAGE COST PER UNIT TIME WHEN PREVENTIVE MAINTENANCE IS PERFORMED EVERY i^{th} INTERVAL BY REPLACING EQUIPMENTS IN STATE a_6

Replacement Interval	Cost Equation	Cost Schedules* 1: $k_1 = 11$	[illegible]: $k_1 = 15$	[illegible]: $k_1 = 19$
1	$\frac{1}{1}$ [$0k_1$ + $.143k_2$ + k_3]	1.57	1.57	1.57
2	$\frac{1}{2}$ [[$.132k_1$ + $.132k_2$ + k_3]	1.49	1.[illegible]	2.0[illegible]
4	$\frac{1}{4}$ [.3[illegible]k_1 + .1[illegible]2k_2 + k_3]	1.42	1.[illegible]0	2.1[illegible]
6	$\frac{1}{6}$ [.60[illegible]k_1 + .1[illegible]k_2 + k_3]	1.[illegible]8	1.[illegible]9	2.20
8	$\frac{1}{8}$ [$.767k_1$ + .1[illegible]0k_2 + k_3]	1.24	1.6[illegible]	2.01
10	$\frac{1}{10}$ [[1.1[illegible]k_1 + .1[illegible]k_2 + k_3]	1.3[illegible]	1.[illegible]	2.[illegible]

* In each of the three cost schedules, k_2 = [illegible] and k_3 = 1.

TABLE 5

GROUP II COST EQUATION: AVERAGE COST PER UNIT TIME WHEN PREVENTIVE MAINTENANCE IS PERFORMED EVERY i^{th} INTERVAL BY REPLACING EQUIPMENTS IN STATES a_5 AND a_6

Replacement Interval	Cost Equation	Cost Schedules* 1: $k_1 = 8$	[illegible]: $k_1 = 11$	[illegible]: $k_1 = 15$	[illegible]: $k_1 = 19$
1	$\frac{1}{1}$ [$0k_1$ + $.2k_2$ + k_3]	1.80	1.80	1.80	1.80
2	$\frac{1}{2}$ [$0k_1$ + $.3750k_2$ + k_3]	1.25	1.25	1.[illegible]	1.[illegible]
4	$\frac{1}{4}$ [$.158k_1$ + $.507k_2$ + k_3]	1.07	1.19	1.35	1.51
6	$\frac{1}{6}$ [$.332k_1$ + $.565k_2$ + k_3]	.986	1.15	1.[illegible]	1.59
8	$\frac{1}{8}$ [$.789k_1$ + $.277k_2$ + k_3]	1.05	1.35	1.74	2.14
10	$\frac{1}{10}$ [$.987k_1$ + $.356k_2$ + k_3]	1.03	1.33	1.[illegible]2	2.12

* In each of the four cost schedules, k_2 = 4 and k_3 = 1.

$$C = [.1233k_1 + 0k_2 + 0k_3]$$

Schedule 1 = 0.9

Schedule 2 = 1.36

Schedule 3 = 1.85

Schedule 4 = 2.34

The curves in Figures 6 and 7 show the effect of different cost schedules and replacement times on the average hourly cost of replacing, respectively, state a_6 equipments only and state a_6 and state a_5 equipments. Figure 6 indicates that preventive maintenance is consistently less costly than repair of in-service failure when the latter has values of $k_1 = 15$ and $k_1 = 19$ (see two top curves). When repairs of in-service failures have values of $k_1 = 11$, preventive maintenance is less costly only when performed somewhere between the sixth and the tenth intervals, the optimum time being at the end of the eighth interval.

In Figure 7, showing the variation in costs of replacing equipments in states a_6 and a_5, an extra curve based on schedule 1 costs is included. When this schedule is assumed, the minimum cost occurs when preventive maintenance is performed at the end of the sixth interval; however, at this point, preventive maintenance is still slightly more expensive than no preventive maintenance. For $k_1 = 11$, the optimum point occurs at the end of the sixth interval, and there is a distinct advantage in preventive maintenance except when performed at the end of the first interval. A similar advantage is gained by performing preventive maintenance when $k_1 = 15$ and $k_1 = 19$, but the optimum time for replacement is now the end of the second rather than the end of the sixth interval.

AR-0203

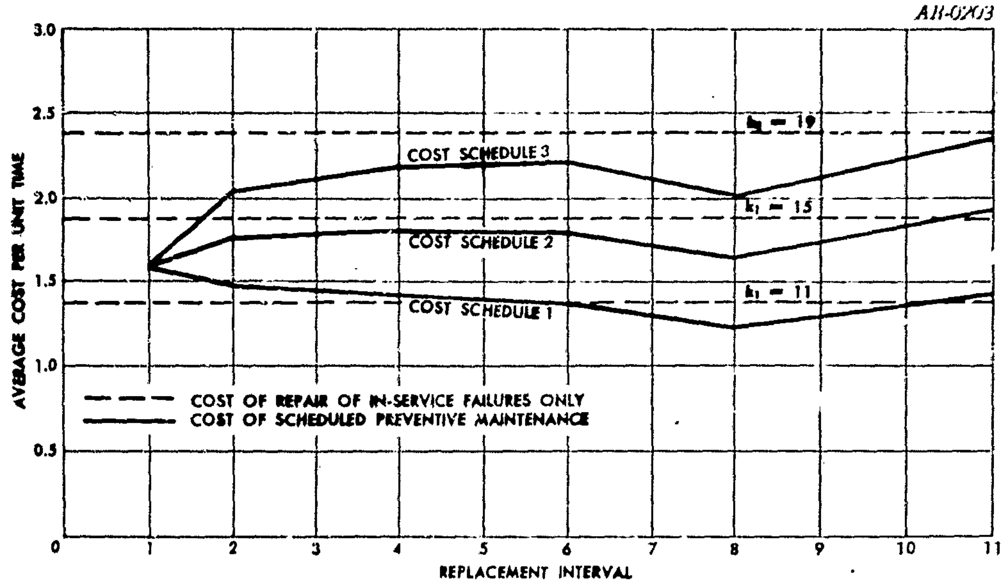

FIGURE 6
COST OF REPAIR OF IN-SERVICE FAILURE vs. COST OF SCHEDULED PREVENTIVE MAINTENANCE, WHEN EQUIPMENTS IN STATE a_6 ARE REPLACED AT VARIOUS INTERVALS

AR-0203

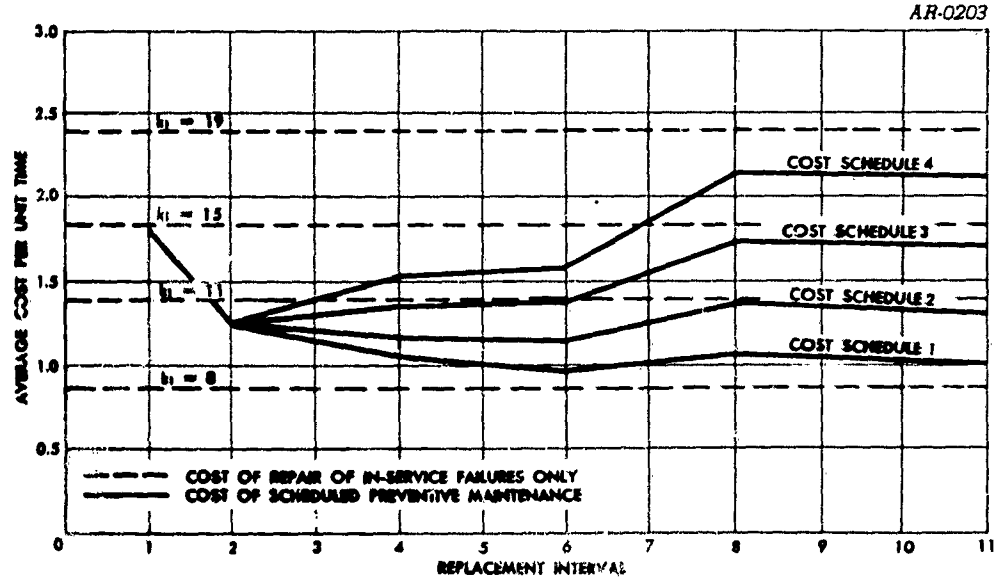

FIGURE 7
COST OF REPAIR OF IN-SERVICE FAILURE vs. COST OF SCHEDULED PREVENTIVE MAINTENANCE, WHEN EQUIPMENTS IN STATES a_6 AND a_5 ARE REPLACED AT VARIOUS INTERVALS

APPENDIX

1. Selection of Maintenance Interval and Time Period for the Transition Matrix

The Markov process requires that the transition matrix be raised to powers higher than one in order to determine the probable states of the equipment after, say, n time intervals. The method assumes that if an equipment fails in service, it will not be returned to service until the beginning of the next time interval.

If the interval selected is one day, for example, it must be assumed that equipments which fail during the day will be repaired during the night shift and be returned to service the next morning. However, suppose that in practice the equipments which fail are replaced in one-half hour's time by the substitution of other equipments. This would mean that second-generation equipments will be introduced within the time period used for the transition matrix. If this should occur, then increasing the powers of the transition matrix, say to n, will not generate the same probable states that actually would result at the end of n periods. However, if the transition matrix is based upon a time interval of one-half hour, then successive chainings of the transition matrix will always match the actual performance pattern. The time interval of the transition matrix must correspond to the repair interval, and the time interval between the times of data collection must be correspondingly short.

Grouping of Data

In the example presented on page 20, states a_1, a_2, a_3 and a_4 were used for illustrative purposes, with no discussion of how they would be chosen in an application of this method. In practice, the raw data would best be recorded as direct numerical observations, especially if the instrumentation were sensitive or the measurement scale were long. Then the definition of states would be the practical problem of data grouping.

Assume that the curve in Figure 8 represents the average change in performance with time, and that t_o and t_n denote, respectively, the beginning and end of the period of observation. The following method is one which might be used to select state boundaries. The time interval, t_o to t_n, may be divided into an arbitrary number of different groupings which are equal to the number of states to be used in the transition matrix. The selection of this number is entirely one of judgment. For example, assume that eight is the selected number of states, and divide the time axis into eight uniform divisions. Vertical and horizontal lines drawn from these divisions to intersect the curve, as indicated in Figure 8, will divide the vertical scale into as many different classes as there are divisions of the time axis. The probability that an equipment will be in any one of these groupings is computed by the method illustrated. These groupings will not produce uniform, numerical divisions on the ordinate scale unless the average curve is linear. Nevertheless, the divisions will be "uniform" insofar as they represent an average range of performance covered by the equipment in the uniform time period selected on the time axis.

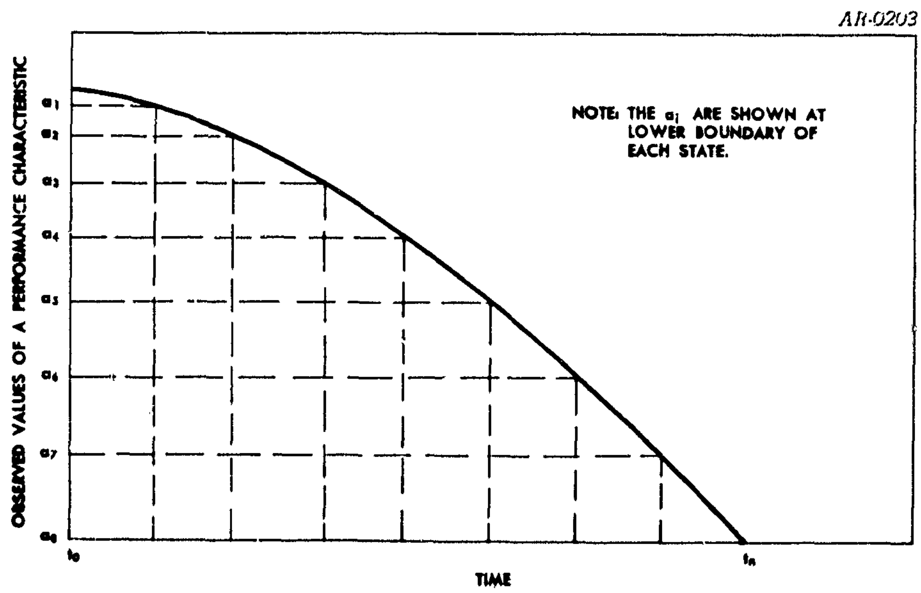

FIGURE 8
SELECTION OF STATE BOUNDARIES FROM OBSERVED VALUES OF A PERFORMANCE CHARACTERISTIC

There is one further consideration. The average curve may be nearly horizontal for low values of t, as shown in Figure 5 -- for example, when parallel redundancy exists. In this instance, the performance measurement will tell nothing of the probable relationship of the equipment to the time axis in this interval. However, the transition matrix can reflect the fact that the equipment is moving uniformly through this time interval or through the classes below the horizontal portion of the curve. Such a matrix would have zero's and one's in the columns which pertain to these classes as in matrix B, page 39. It should be noted that the performance characteristic measured will usually be an operating characteristic correlated with, but not identical to, the part deterioration characteristic itself.

If the foregoing procedure should result in the inclusion of certain "unnecessary" states, these should be eliminated. For example, if the transition matrix should indicate very low probabil- ' .es that an equipment will be in state a_k at the beginning of the interval, it is proper to say that state a_k is unimportant and can be combined with state a_{k-1} or state a_{k+1}.

3. Method for Computation of p_{ij} of Transition Matrix from Empirical Data

To illustrate the method for computing the p_{ij} of the transition matrix from empirical data, assume four equipments and four levels of performance. They are observed for a maximum of eight time intervals. The data -- which could be derived just as well from observations on one equipment which was repaired or returned to operating level a_1 four times -- are listed in Table 6. The point to stress is that each line represents the deterioration of one equipment from a higher operating state to failure, state a_4. If the equipment fails and is restored to a higher state through repair, another line must be added to the table.

In recording data for use in developing a transition matrix, it is not sufficient simply to list all equipments which are in a given state at the end of any given time period, for this information does not indicate the states which the equipments were in before progressing to the observed states. The observed state and the immediately preceding state must both be recorded for each equipment at the end of each interval, so that the matrix will indicate the transition from one state to another within the interval.

TABLE 6 FREQUENCY OF PASSAGE FROM ONE GIVEN STATE TO ANOTHER IN A GIVEN TIME INTERVAL, FOR FOUR EQUIPMENTS																		
Equipment Number	Part I State in Each Interval									Part II Frequency of Passage from State a_i to State a_j in One Interval								
	0	1	2	3	4	5	6	7	8	a_{11} (1)	a_{12} (2)	a_{13} (3)	a_{14} (4)	a_{22} (5)	a_{23} (6)	a_{24} (7)	a_{33} (8)	a_{34} (9)
1	a_1	a_1	a_1	a_1	a_2	a_2	a_2	a_3	a_4	3	1			2	1			1
2	a_1	a_1	a_1	a_3	a_3	a_3	a_3	a_3	a_4	2		1					4	1
3	a_1	a_1	a_1	a_1	a_1	a_4				4			1					
4	a_2	a_2	a_2	a_2	a_2	a_2	a_3	a_3	a_4					5	1		1	1
Total										9	1	1	1	[illegible]	[illegible]		5	3

Part I of the table lists the states observed for each of the equipments in each of the eight intervals. The values at time 0 are beginning values, those under "1" are values at the end of the first interval, and so on. Thus, Equipment No. 3 was observed to be in state a_1 for four intervals, and it failed during the fifth interval. Part II indicates the number of times the equipments passed from one given state to another in one interval. The first four columns give the frequency of passage from state a_1 to lower states; the next three columns give the frequency of passage from state a_2 to lower states; and the last two columns give the frequency of passage from state a_3 to failure, a_4. The p_{ij} are computed from these frequencies.

The following tabulation and estimated stochastic matrix are determined from the values in Table 6.

$$
\begin{array}{c|cccc|c}
 & a_1 & a_2 & a_3 & a_4 & \Sigma \\
a_1 & 9 & 1 & 1 & 1 & 12 \\
a_2 & & 7 & 2 & 0 & 9 \\
a_3 & & & 5 & 3 & 8
\end{array} .
$$

$$
\begin{array}{c}
 \\
a_1 \\
a_2 \\
a_3 \\
a_4
\end{array}
\begin{array}{c}
\begin{array}{cccc}
a_1 & a_2 & a_3 & a_4
\end{array} \\
\left\| \begin{array}{cccc}
9/12 & 1/12 & 1/12 & 1/12 \\
 & 7/9 & 2/9 & 0 \\
 & & 5/8 & 3/8 \\
 & & & (1)
\end{array} \right\|
\end{array} .
$$

4. <u>Failure Density and Reliability Functions</u>

The unreliability function is developed by matrix multiplication if one of the states is defined as failure. In the notation used throughout this paper, the state in the right-hand position in the matrix has been so defined. (In the examples, this has been state a_4 or state a_7.) The density function is obtained by computing the differences between successive values of the unreliability function. As an example, take the four-state transition matrix given previously,

$$
\left\| \begin{array}{cccc}
.5 & .5 & 0 & 0 \\
0 & .5 & .5 & 0 \\
0 & 0 & .1 & .9 \\
0 & 0 & 0 & 1
\end{array} \right\| .
$$

The distribution by state, the values of the unreliability function, and the values of the density function for integral values of time from $t = 0$ to $t = 11$ are shown in Table 7, together with the density value for $t > 11$. Table 7 shows unrounded values in all cases.

Before illustrating how the density function $u(t)$ can be used to compute the mean and variance, it is of interest to show how the mean time-to-failure is computed directly from the transition matrix. Denote by z_i the mean time-to-failure of an equipment, given that the performance level is in state i. Then z_1 is the mean time-to-failure of a new equipment -- i.e., an equipment with performance

TABLE 7*

COMPUTATIONS FOR THE NUMERICAL EXAMPLE

Time (t)	State				Failure-Density Function	Reliability Function	Average State Values
	a_1	a_2	a_3	a_4 Failure U(t)	u(t)	R(t)	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0	1	0	0	0		1	1.00
1	.5	.5	0	0	0	1	1.50
2	.25	.5	.25	0	0	1	2.00
3	.125	.375	.275	.225	.225	.775	2.60
4	.0625	.250	.215	.4725	.2475	.5275	3.10
5	.03125	.15626	.1465	.6660	.1935	.3340	3.45
6	.015625	.09375	.092775	.79785	.13185	.20215	3.67
7	.0078125	.0546875	.0561525	.8813475	.0384675	.1186425	3.81
8	.00390625	.03125	.032959	.93188475	.05053725	.06811525	3.89
9	.001953125	.017578125	.0189209	.96154785	.02966310	.03845215	3.94
10	.0009765625	.009765625	.0106811525	.97857666	.01702831	.02142334	3.97
11	.00048828125	.00537109375	.00595092775	.98818969725	.00961303725	.00181030275	3.98
t > 11	0	0	0	1.000	.01181030275	0	4.00

* Originally presented as Table 1, this table is repeated here for the convenience of the reader.

in the highest level, state a_1. Since state a_4 denotes failure in this example, $z_4 = 0$.

The mean times-to-failure satisfy the following system of linear equations:

$$z_1 = .5\ (z_1 + 1) + .5\ (z_2 + 1)$$

$$z_2 = .5\ (z_2 + 1) + .5\ (z_3 + 1)$$

$$z_3 = .1\ (z_3 + 1) + .9.$$

The solution of this system of equations is

$$z_3 = 10/9,\ z_2 = 28/9, \text{ and } z_1 = 46/9.$$

The justification of each of the foregoing equations is dependent upon an argument of the following type. Consider the first equation of the system as an example. If an equipment is in state a_1 initially, its mean life is z_1, the left side of the equation. The right side of the equation expresses this mean life as a two-step evaluation: the transition in one time interval and the expected mean life thereafter. If an equipment is in state a_1, the best estimate of its expected life remains z_1 in the absence of other information regarding this variable. Thus, the .5 of the equipments that remain in state a_1 at the end of the time interval have an expected life of z_1 to be added to the one time interval previously survived. The .5 of the equipments which deteriorate to state a_2 have, at the end of the time interval, an additional expected life of z_2. The sum of these two gives the right side of the first equation of the set. The second and third equations are justified in the same manner. It should be noted that the third equation simplifies, since $z_4 = 0$.

When the mean and variance are computed directly from the density function (which here has the form of an open-end distribution) it is necessary, or at least advisable, to include the failures which occur after the 11 time intervals shown in the computation. This can be done by selecting an approximate length of life for these failures which will yield the correct mean for the failure-density function, u(t), as computed above, $z_1 = 46/9$. In the example, this turns out to be t = 13.2. Thus, it is assumed as an approximation that at t = 13.2, there were .01181030275 failures. It is now possible to use ordinary formulas to compute the mean life (5.111) and the variance (4.0972).

It should be noted that the preceding computation is based on the assumption that failures occur exactly at the end of the time interval. An alternative assumption is that failures occur at the midpoint. Using this assumption, the mean would be reduced to 4.611 and the variance would be unchanged.